1998

Examiner Errors in Scoring the WISC-III Vocabulary, Comprehension, and Similarities Subtests

Ryan L. Neitzel
Eastern Illinois University

THESIS REPRODUCTION CERTIFICATE

TO: Graduate Degree Candidates (who have written formal theses)

SUBJECT: Permission to Reproduce Theses

The University Library is receiving a number of request from other institutions asking permission to reproduce dissertations for inclusion in their library holdings. Although no copyright laws are involved, we feel that professional courtesy demands that permission be obtained from the author before we allow these to be copied.

PLEASE SIGN ONE OF THE FOLLOWING STATEMENTS:

Booth Library of Eastern Illinois University has my permission to lend my thesis to a reputable college or university or the purpose of copying it for inclusion in that institution's library or research holdings.

7/22/98
Date

I respectfully request Booth Library of Eastern Illinois University **NOT** allow my thesis to be reproduced because:

_____

_____

_____

_____ _____

Author's Signature Date

Examiner Errors in Scoring the WISC-III Vocabulary, Comprehension, and Similarities Subtests

(TITLE)

BY

Ryan L. Neitzel

1972 -

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

Specialist in School Psychology

IN THE GRADUATE SCHOOL, EASTERN ILLINOIS UNIVERSITY
CHARLESTON, ILLINOIS

1998

YEAR

I HEREBY RECOMMEND THIS THESIS BE ACCEPTED AS FULFILLING
THIS PART OF THE GRADUATE DEGREE CITED ABOVE

22 July 1998

DATE

22 July 1998

DATE

Abstract

Examiner errors when scoring the WISC-III Vocabulary, Comprehension, and Similarities subtests were examined. Twenty-one graduate students enrolled in school psychology training programs and twenty-three certified school psychologist practitioners participated in scoring a constructed record form containing both standard responses (items clearly accounted for in the WISC-III *Manual*) and ambiguous responses (items not specifically accounted for in the WISC-III *Manual*). Descriptive Statistics demonstrated that graduate students and practitioners made errors on all subtests and item types. However, the Vocabulary, Comprehension, and Similarities ambiguous responses contained the most errors. Furthermore, Analysis of Variance demonstrated that for each subtest, ambiguous responses were responsible for greater errors than standard responses and that graduate students and practitioners do not differ when scoring the same WISC-III record form. Results suggest that the WISC-III has made improvements over its predecessors in assisting examiners when scoring actual items on the Vocabulary, Comprehension, and Similarities subtests.

Acknowledgements

I would like to thank Dr. Gary L. Canivez, Chair of my Thesis Committee, for his guidance and encouragement through every stage of this thesis. Without his assistance, completion would not have been possible. Additionally, I would like to recognize Dr. Kevin Jones and Dr. Joe Williams for agreeing to serve on the committee and for their comments and suggestions during the completion of this project. Special thanks are given to Dr. Marley Watkins of Pennsylvania State University for his assistance in data collection. Lastly, I want to give thanks to the participants who were responsible for providing the necessary data in order to finalize this study.

Table of Contents

List of Tables

Chapter 1

Examiner Errors in Scoring the WISC-III Vocabulary, Comprehension, and Similarities Subtests

Although several intelligence measures are available for use in psychological practice, the Wechsler Intelligence Scales are the most commonly administered individual intelligence tests for both adults and children (Oakland & Zimmerman, 1986; Piotrowski & Keller, 1989). Furthermore, because these scales are the most commonly administered individual intelligence tests, they are also the most frequently taught intelligence tests in school psychology training programs (Oakland & Zimmerman, 1986). Because of their frequent use, it is imperative that examiners be qualified and competent when utilizing these instruments. Without competence, it is clear that the obtained results would not be representative of the examinee's true abilities and any decisions would indeed be questionable.

While previous research has shown that both graduate students and practitioners make frequent errors when administering and scoring the Wechsler Intelligence Scale for Children-Revised (i.e., Slate & Jones, 1990) and the Wechsler Adult Intelligence Scale-Revised (i.e., Slate, Jones, Murray, & Coulter, 1993), research regarding examiner scoring errors on the Wechsler Intelligence Scale for Children-Third Edition is lacking. Furthermore, research examining scoring differences between practitioners and graduate students is also lacking. These errors are important because they directly impact the accuracy of obtained test scores. Accuracy of intelligence test scores is important because

these scores can have a significant effect on children's lives, especially special education decisions. Accordingly, there is a need to conduct research that carefully addresses scoring errors and differences on the Wechsler Intelligence Scale for Children-Third Edition (WISC-III). Specifically, research is needed regarding the conditions responsible for scoring differences on subtests that appear particularly prone to errors (i.e., Vocabulary, Similarities, and Comprehension).

Because the WISC-III is similar to its previous versions (WISC, WISC-R) in many areas (i.e., reliability coefficients), it may also be subject to a large number of scoring errors. However, differences in test design do indeed exist (i.e., differing subtest items, expanded scoring examples). These differences bring into question whether or not previous research generalizes to the WISC-III. By conducting such research on the WISC-III, practitioners and graduate students can become aware of potential scoring difficulties that may contribute to the reduced reliability and possible invalidity of WISC-III test results.

With research investigating examiner scoring errors on the WISC-III lacking, the present study was conducted to answer the following research questions: (1) Do graduate students and practitioners make errors in scoring the WISC-III? (2) If so, are certain subtest items (standard vs. ambiguous) more prone to error than others? (3) Do graduate students and practitioners differ in their scoring of the same WISC-III record form?

Chapter 2

Review of Literature

Sattler, Winget, and Roth (1969) examined scoring difficulty of Wechsler Adult Intelligence Scale (WAIS) and Wechsler Intelligence Scale for Children (WISC) Comprehension, Similarities, and Vocabulary subtests. Ambiguous responses selected from actual test protocols or constructed by the authors were scored by eight doctoral level clinical psychologists for the WAIS. Ambiguous WISC responses were scored by a group of eight doctoral clinical psychologists, five of whom had also scored the WAIS protocols (Sattler et al., 1969). Overall, unanimous scoring agreement was found for only twelve percent of the WAIS responses and only eight percent of the WISC responses. These results clearly demonstrated that examiners disagreed on scoring ambiguous WAIS and WISC responses (Sattler et al., 1969). Furthermore, these differences contributed to reduction in test reliability. It was recommended that more thorough scoring standards could enhance test reliability and accuracy in IQ score obtained.

Miller, Chansky, and Gredler (1970) investigated the degree of agreement among school and clinical psychologists in training for the scoring of WISC protocols. A total of 24 school psychology and 8 clinical psychology trainees were asked to score a WISC protocol, developed by the authors, which contained a sample of responses not covered in the WISC *Manual* (Miller et al., 1970). Results showed that the raters did not demonstrate perfect agreement on any of the subtests and, with the exception of the Digit Span subtest, the range of scaled score units exceed the standard error of measurement for each of the subtests at the three age ranges reported by Wechsler (Miller et al., 1970).

For example, scaled scores on the Comprehension subtest ranged from 4 to 11. This low rater agreement appeared related to the difficulty in scoring items not clearly accounted for in the *Manual*, failure to follow guidelines in relation to cut-off criteria, and a failure by raters to examine all responses (Miller et al., 1970). Although trainees were involved, this study indicated that scoring errors cannot be tolerated and raises the question of how widespread these errors might be occurring.

In a study designed to investigate the degree of agreement among professional psychologists in scoring WISC Record Forms, Miller and Chansky (1972) examined a fabricated Record Form independently scored by sixty-four psychologists and found a wide range of scaled scores for each subtest. Additionally, it was noted that the Information, Vocabulary, Similarities, and Comprehension subtests appeared to produce the greatest interscorer variability (Miller & Chansky, 1972). Furthermore, Full Scale IQs also varied. Miller and Chansky (1972) reported that Full Scale IQ on the same Record Form ranged from 78 to 95. The two reasons given for these results were that responses were not specifically provided in the *Manual* and failing to check all responses carefully (Miller & Chansky, 1972). Overall, this study demonstrated that scoring differences among professional examiners do occur and that these differences can have a serious impact on the Full Scale IQ obtained.

Warren and Brown, Jr. (1973) examined 10 types of examiner scoring errors on 120 WISC and 120 Stanford-Binet protocols obtained from 40 graduate students in four IQ measurement classes taught by four different instructors. Scoring errors included failure to record a response, failure to follow procedures specified in the *Manual*, scoring,

and tabulating. A total of 1,939 errors were found in the 1,873 subtests examined, with 725 occurring on WISC protocols (Warren & Brown, Jr., 1973). Furthermore, 37 percent of the protocols contained errors that affected the reported IQ, with a change of 1 to 16 points occurring on WISC protocols (Warren & Brown, Jr., 1973). Results from this study emphasized the need for examiners to be more aware of potential scoring errors and how to avoid these problems.

Oakland, Lee, and Axelrad (1975) conducted a study examining experienced psychologists scoring of WISC protocols to determine the degree of disagreement. Results showed that differences among the 94 examiners in scoring actual protocols tended to be within an acceptable range as established by the standard error of measurement (Oakland, et al., 1975). However, it should be noted that this study did not reflect individual subtest items that were responsible for the overall Verbal, Performance, and Full Scale IQ scores. It was also reported that on several occasions the Vocabulary subtest exceeded its' corresponding standard error of measurement (Oakland et al., 1975). While results of this study suggest that examiner differences in scoring WISC protocols are within acceptable levels, a need to determine potential errors is still warranted.

Sherrets, Gard, and Langner (1979) investigated the frequency of clerical errors appearing on WISC protocols. A total of 200 protocols were selected from patient files: 100 from a psychiatric facility and 100 from public school psychological services records. Results showed that 46.5 percent of the 200 protocols scored by 39 examiners from 17 psychiatric facilities or 22 school systems contained one or more errors and that these errors were responsible for as much as a nine point increase or seven point decrease

in Full Scale IQ. Clearly, this is an extremely high percentage of errors and suggests serious problems with reliability and validity (Sherrets et al., 1979).

Bradley, Hanna, and Lucas (1980) completed a study designed to examine scorer differences on two separate Wechsler Intelligence Scale for Children-Revised (WISC-R) protocols that were developed by the authors. A total of 63 National Association of School Psychologists (NASP) members independently scored both protocols. The first record form contained no ambiguous responses and no administration errors, while the second record form contained ambiguous responses, administrative errors related to beginning and discontinuation points in the subtests, too much questioning of some examinee responses, and a few highly unusual responses (Bradley et al., 1980). Results demonstrated that IQ scores could easily vary by six to eight points based on standard deviations. Bradley et al. (1980) emphasized that their overriding impression concerning reliability of scoring was that no WISC-R is immune to serious scoring errors and that users of such tests should redouble their efforts to maximize reliability of scoring by rigorous adherence to standardized scoring procedures.

Conner and Woodall (1983) conducted a study to determine the effects of experience in the administration and scoring of the Wechsler Intelligence Scale for Children-Revised (WISC-R). Ten graduate students administered 15 WISC-Rs and scored the record forms. Nine record forms from each student were randomly drawn and evaluated for four types of errors (Response Scoring, IQ, Administrative, and Mathematical) and for the Total Error rate. Response Scoring errors involved assigning values to responses other than specified in the *Manual*, while Mathematical errors

included such things as miscalculating chronological age and inaccurate summation of individual subtests. Additionally, IQ errors involved items such as incorrect conversion of raw scores to scale scores and inaccurate calculation of Verbal, Performance, and Full Scale IQ's, while Administrative errors included such things as failing to probe as indicated in the WISC-R *Manual*, failure to record verbal responses and failure to credit items not administered below the established subtest basal. Total Error rate was determined by combining all errors in each of the four types. It was determined that of the five types of errors only the Total Errors made and Administrative errors made were significantly (.001) decreased with experience and structured feedback (Conner & Woodall, 1983). This suggests that even with experience and specific training on the WISC-R, examiners continued to make errors that affected the reliability and validity of the WISC-R results. Furthermore, Response Scoring errors should be of most concern to students and practitioners. These errors were more prevalent than all other types combined. The authors concluded by suggesting that with experience, examiners develop individual scoring patterns which may differ from what is required by the WISC-R *Manual* (Conner & Woodall, 1983).

In a study conducted to determine which examinee responses are frequently misscored by graduate student examiners on the WISC-R and WAIS-R, Slate and Jones (1988) analyzed a total of 309 WISC-R protocols and 326 WAIS-R protocols completed by 40 graduate students. To identify scoring errors, a conservative coding procedure was used in which only items clearly misscored according to the Wechsler *Manual* were coded as errors while the scoring of ambiguous responses were regarded as correct (Slate

& Jones, 1988). Results were consistent with previous findings. Graduate students most frequently misscored examinee responses on the Vocabulary, Similarities, and Comprehension subtests and committed scoring errors on items from the Information and Picture Completion subtest (Slate & Jones, 1988). An interesting finding from this study was that for several responses required to be queried by examiners, it was not uncommon for the examiner to simply assign a higher point value to the response rather than question as required (Slate & Jones, 1988). It seems evident that this study supports the need for more clearly defined scoring criteria in the *Manuals* and a strong need for examiners to strictly follow the scoring criteria as outlined in the *Manuals*.

Slate and Chick (1989) examined fourteen graduate students WISC-R record forms to determine the frequency and types of errors made on the WISC-R. Two types of errors were examined. Independent errors included mechanical errors, scoring errors, errors in questioning, errors in determining basal and/or ceiling, and errors in converting raw scores to scaled scores. Total errors included independent errors combined with the resulting changes in raw scores, standard scores, and IQ scores (Slate & Chick, 1989). On average, students committed 8.1 independent errors and 15.2 total errors on each record form. Vocabulary, Comprehension, and Similarities were the three subtests on which students made the most mistakes (Slate & Chick, 1989). Incorrect point assignment was the most frequent error reported and student examiners were three times as likely to award more points for examinee responses than permitted by the WISC-R *Manual* as they were to award fewer points (Slate & Chick, 1989). This may result in inflated Verbal IQ and Full Scale IQ scores and possibly influence placement decisions. The authors

concluded by stating a need for greater clarity in what is acceptable as a 2, 1, or 0 point response is warranted and that psychologists in the field may be making diagnosis and placement decisions based on inaccurate information (Slate & Chick, 1989).

Slate and Jones (1990) conducted a study to investigate the most frequent types of examiner errors made by graduate students in administering the WISC-R and on which items mistakes were most likely to occur. A total of 26 participants were randomly assigned to administer the WISC-R either five or ten times to volunteer examinees. All students were enrolled in an individual intelligence testing course (Slate & Jones, 1990). Results showed that students averaged 11.3 errors on each WISC-R protocol and none of the 217 Record Forms were without error (Slate & Jones, 1990). Furthermore, when the errors were corrected, Full Scale IQ scores were changed on 79.7 percent of the Record Forms (Slate & Jones, 1990). Although analysis of errors determined that the most frequent error was a failure to record the examinee's responses, which may not impact the obtained score, it was found that the second most frequent error was the examiner assigning incorrect point values to examinee responses. Clearly, assignment of incorrect point value has the potential to greatly effect the obtained score. Overall, this study supported the need for more explicit criteria in order to facilitate more accurate scoring of the WISC-R.

Slate, Jones, Coulter, and Covert (1992) studied practitioner's administration and scoring of the WISC-R to determine whether practitioners make errors in scoring and, if so, what types are made. Results demonstrated that practitioners committed errors on all 56 WISC-R protocols randomly selected from school psychological records, with the

most mistakes occurring on the Vocabulary subtest. Further analysis showed that errors were frequently made in the form of failing to record a response and assigning incorrect point values. These errors were responsible for as much as a four point deviation from the correct Full Scale IQ (Slate et al., 1992). They also suggested that through better preservice and inservice training these errors could be corrected and significantly reduced.

Franklin, Jr., Stillman, Burpeau, and Young (1982) examined the extent of examiner error during administration of the WAIS by practicing school psychologists and school psychology graduate students eligible for state certification as psychometrists. Each examiner administered the WAIS to one of four clients who had been trained to give standard verbatim responses (Franklin, Jr. et al., 1982). Examiner obtained scores were then compared to "true" scores calculated with 100% agreement between two practicing school psychologists. Results showed that a large number of errors directly affecting obtained scaled scores on Information, Comprehension, and Vocabulary subtests were committed by the examiners (Franklin, Jr. et al., 1982). For example, of the 13 examiners who administered the WAIS to the second client, 12 (92.3%) made errors resulting in the Vocabulary subtest having a different scaled score than the "true" score. In addition to the previously mentioned subtests, errors were found on all remaining subtests regardless of subjective or objective scoring criteria. The authors proposed that because of these errors there is a need for reexamination of training procedures at the university level to provide more stringent checks and feedback to students learning how to administer and score intelligence tests. They also recommended continuing education

for school psychologists to facilitate proficiency with new and revised psychometric instruments (Franklin, Jr. et al., 1982).

Slate and Jones (1990) conducted a study to investigate specific problems caused by the traditional method of teaching students to administer the Wechsler Adult Intelligence Scale-Revised (WAIS-R). Analysis of 180 Record Forms completed by 26 graduate students revealed an average of 8.8 mistakes per Record Form, with 177 of them containing errors (Slate & Jones, 1990). After correcting these errors, 81% of Full Scale IQ's were changed, with 62.2% being lower than the IQ assigned by the students and 17.8% being higher (Slate & Jones, 1990). Furthermore, corrected IQ's were different from assigned IQ's by more than two points on 64% of the Record Forms, more than three points on 16% of the Record Forms, with a maximum deviation of seven points (Slate & Chick, 1990). As found in previous studies, Vocabulary, Comprehension, and Similarities were the subtests which students made the most mistakes. Furthermore, the most frequent error was not recording an examinee's response verbatim. The second most frequent error was assigning incorrect point value to a response (Slate & Jones, 1990). These results also indicated a need for more clearly specified scoring criteria and instruction that focuses on these difficult to score subtests.

In research designed to investigate the frequency and types of errors graduate student examiners make on actual WAIS-R Record Forms, Slate and Jones (1990) analyzed 149 record forms completed by 22 masters level students enrolled in an individual intelligence testing course. To strengthen the accuracy of results, a protocol entry was coded as an error only when it clearly violated the test *Manual* instructions

(Slate & Jones, 1990). Results supported previous findings. Students made an average of 7.95 errors per protocol, with 145 protocols (97.3%) containing at least one error (Slate & Jones, 1990). Additionally, most errors occurred on the Vocabulary, Comprehension, and Similarities subtests (Slate & Jones, 1990). Based on these findings, it is clear that examiner error decreases the reliability and validity of obtained scores on the WAIS-R and other Wechsler scales.

Slate, Jones, Murray, and Coulter (1993) conducted a study to determine if practitioners committed errors in administering and scoring the WAIS-R. A total of 50 Record Forms obtained from eight practitioners were analyzed for errors, with items clearly indicated as incorrect in the WAIS-R *Manual* being the only responses considered (Slate et al., 1993). Results of this study showed that practitioners committed errors on all 50 Record Forms, with the mean number of errors for individual practitioners ranging from 13.4 to 103.8 per Record Form (Slate et al., 1993). Additional analyses determined that 54% of the Record Forms were in need of a corrected IQ score (Slate et al., 1993). Of these, 23 were lower than those assigned by the practitioner and four were higher (Slate et al., 1993). Furthermore, Slate et al. (1993) noted that deviations in Full Scale IQ of as much as five points were not uncommon. Interestingly, this study compared the present results with errors made by graduate students in a previous study (Slate & Jones, 1990) and found that practitioners made almost twice as many errors as the graduate students. These results suggested that professionals may be more prone to error than graduate students and urged practitioners to strictly check their Record Forms several times for errors and refer to the WAIS-R *Manual* when scoring rather than relying on

memory (Slate et al., 1993).

In a study designed to investigate potential error in administering and scoring the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R), Whitten, Slate, Jones, Shine, and Raggio (1994) analyzed 57 record forms completed by two doctoral level interns, two post-master's level doctoral students, and three doctoral level practitioners who were Nationally Certified School Psychologists. Examiners on all 57 record forms committed errors. Failing to record a response and assigning too many points to an examinee response were the most frequent errors (Whitten et al., 1994). When examiner errors were corrected, 26 record forms had no change in Full Scale IQ, 13 record forms were one to six points higher than the Full Scale IQ assigned by the examiner, and 18 record forms had Full Scale IQ's one to fourteen points lower than the Full Scale IQ assigned by the examiner (Whitten et al., 1994). Whitten et al. (1994) emphasized how these results are consistent with completed research on the WISC-R and WAIS-R and suggested that carelessness and inadequate training may be responsible for the errors that occurred.

Based on these studies, it is apparent that examiners do make errors when scoring the Wechsler intelligence tests (i.e., WISC, WISC-R, WAIS, WAIS-R, WPPSI-R). These errors include, but are not limited to incorrect age assignment, failure to follow established discontinue criteria, inaccurate conversion of raw scores to scaled scores, and incorrect assignment of point values to a given response. However, research examining scoring errors on the WISC-III is lacking. Due to this, it is important that research be conducted in order to determine whether or not examiners continue to make previously

identified errors, new errors, or no errors at all.

Additionally, while previous research has examined scoring errors by graduate students (Slate & Jones, 1988; Slate & Chick, 1989; Slate & Jones, 1990) and practicing psychologists (Miller & Chansky, 1972; Bradley, Hanna, & Lucas, 1980) as well as standard items (Slate & Jones, 1988) and ambiguous items (Sattler, Winget, & Roth, 1969), research has yet to be conducted examining examiner type and item type simultaneously. Accordingly, it would be beneficial to conduct research in order to investigate the interaction between scorer type (Graduate Student/Practitioner) and item type (Standard/Ambiguous).

Chapter 3

Method

Participants

Twenty-one students enrolled in either Specialist level (n=17, 81%) or Doctoral level (n=4, 19%) School Psychology training programs were solicited by mail to participate. Nine School Psychology training programs were contacted requesting student participation. Of these nine programs, students from three submitted data. Student participants reported completing from one to four Intellectual Assessment courses.

Forty-seven certified school psychologist practitioners employed within various school districts and educational cooperatives were solicited by mail to participate. Of these, 23 submitted data for analysis. Practitioner participants reported having between one and eight Intellectual Assessment courses. Additionally, practitioner characteristics closely approximated the national figures presented by Reschly and Wilson (1997). Table 1 presents the percentages of time spent in various school psychological roles by the current sample and Reschly and Wilson's (1997) sample. All participants were treated in accordance with the Ethical Principles of Psychologists and Code of Conduct (American Psychological Association, 1992) and additionally were given an opportunity to receive a cash award by random drawing for their participation.

Materials

The Vocabulary, Comprehension, and Similarities subtest record forms of the Wechsler Intelligence Scale for Children-Third Edition (WISC-III) were utilized. A WISC-III record form was constructed containing responses to each item of the

Vocabulary, Comprehension, and Similarities subtests. Items from the record form were randomly selected as standard or ambiguous, which resulted in 11 Standard and 8 Ambiguous responses for the Similarities subtest, 9 Standard and 9 Ambiguous responses for the Comprehension subtest, and 17 Standard and 13 Ambiguous responses for the Vocabulary subtest.. Standard responses were directly selected from the scoring examples provided in the WISC-III *Manual*, while ambiguous responses were obtained from actual WISC-III administrations and not clearly identified in the WISC-III *Manual*. Ambiguous responses were selected from record forms by children of various ages. Two professors in a School Psychology training program independently scored each item response. Interrater agreement on item scores was 100% and scores awarded served as the scoring key. In addition to the subtest record forms, a demographic information form including the number of Intellectual Assessment courses taken and how practitioners spent their time was utilized (see Appendix A).

Procedure

Participants were provided the set of the three subtests and instructed to score each response as if it were an actual case (see instructions in Appendix B). The WISC-III *Manual* was to be used in scoring as this is standard and required practice. Participants were instructed to score each item regardless if a discontinue rule was met. Participants were asked to return the scored materials in the enclosed self-addressed stamped envelope. The demographic information form was included with each set of materials and contained an identification number to maintain anonymity. WISC-III Record Forms were included for analysis only if each response on the subtests was scored.

Analysis

In order to answer the first research question (Do graduate students and practitioners make errors in scoring the WISC-III?), scoring errors and the effects were examined separately by subtest (Similarities, Vocabulary, and Comprehension). Items were judged correctly scored if they matched the agreed upon score established by the School Psychology training professors. Incorrectly scored items resulted in an error score of one regardless of how large the error may have been. All subtests were scored separately for both standard and ambiguous responses. Descriptive statistics and frequency of examiner errors by subtest item type were determined.

To answer research questions two (Do ambiguous responses result in greater scoring errors?) and three (do school psychology students make greater errors than practitioners), a 2 (Standard or Ambiguous) X 2 (Student or Practitioner) mixed factorial analysis of variance was performed for each WISC-III subtest under investigation (Similarities, Comprehension, and Vocabulary). Response type (Standard or Ambiguous) was the within subjects variable, while the scorer type (Student or Practitioner) was the between subjects variable.

Chapter 4

Results

Table 2 presents the descriptive statistics (means, standard deviations, and ranges) for errors committed by graduate students, practitioners, and the total sample (combined) for each of the three subtests by item type. Similarities, Comprehension, and Vocabulary ambiguous responses contained the most errors. However, examiner errors occurred on all subtests and item types as follows: Similarities Standard (M=.16, SD=.43, 1.46 %), Similarities Ambiguous (M=.80, SD=.88, 10%), Comprehension Standard (M=.27, SD=.59, 3%), Comprehension Ambiguous (M=1.80, SD=1.36, 20%), Vocabulary Standard (M=.39, SD=.78, 2.29%), and Vocabulary Ambiguous (M=3.50, SD=1.68, 26.9%).

Table 3 presents the frequency of examiner errors by subtest item type. Regardless of subtest or item type scoring errors were made. However, participants were more likely to incorrectly score ambiguous items than standard items. The percentage of examiners with perfect scoring agreement ranged from 75 to 86% for Standard responses and from 0 to 47.7% for Ambiguous responses. The Vocabulary subtest was the most problematic because none of the participants were without error for the ambiguous items.

The results of the analysis of variance for the Similarities subtest are presented in Table 4. For this subtest, a significant main effect existed for Item Type where ambiguous items resulted in greater errors than standard items, $F(1,42)=19.19$, $p<.0001$. However, only 17% ($r^2=.17$) of the variability in errors was due to the effect of item type

(Standard vs. Ambiguous). Furthermore, analysis assessing whether scoring differences between graduate students and practitioners existed was not significant, F (1, 42)= 1.55, p=. 220. Lastly, analysis of the interaction effect of Participant Type (Student vs. Practitioner) by Item Type (Standard vs. Ambiguous) have regardless of the level of the other was not significant, F (1, 42)= 2.93, p=. 094.

Table 5 presents the analysis of variance results for the Comprehension subtest. A significant main effect existed for Item Type, F (1,42)=75.05, p<. 0001, in which ambiguous items produced greater errors than standard items. Additionally, 35% (r^2=. 35) of the variability observed in errors was due to the effect of Item Type (Standard vs. Ambiguous). Furthermore, the analysis of whether scoring differences between graduate students and practitioners existed was not significant, F (1,42)=. 06, p=. 808. Lastly, analysis of the effect that Participant Type (Student vs. Practitioner) and Item Type (Standard vs. Ambiguous) have despite the level of the other was not significant, F (1,42)=. 00, p= .995.

Finally, Table 6 presents the analysis of variance results for the Vocabulary subtest. As expected, a significant main effect existed for Item Type, F (1, 42)= 148.42, p<. 0001, where ambiguous items were responsible for more errors than standard items. Additionally, 59% (r^2=. 59) of the variability observed in errors due to the effect of Item Type (Standard vs. Ambiguous). Analysis of scoring differences between graduate students and practitioners was not significant, F (1,42)= .00, p=. 953. The analysis of the effect that Participant Type (Student vs. Practitioner) and Item Type (Standard vs. Ambiguous) have regardless of the other level, F (1,42)=3.53, p=. 067 and indicated that

Item Type did not vary as a function of examiner status. The errors committed were the same for practitioners and students alike.

Chapter 5

Discussion

Although new and revised intelligence measures are becoming available for use in psychological practice, the Wechsler Intelligence Scales continue to be the most commonly administered individual intelligence tests for both adults and children (Oakland & Zimmerman, 1986; Piotrowski & Keller, 1989; Stinnett, Havey, & Ohler-Stinnett, 1994). Due to their frequent use, these scales are given a significant amount of attention in school psychology training programs (Oakland & Zimmerman, 1986). Because of the importance placed on intelligence test results, it is crucial that both graduate students and practicing psychologists be competent when utilizing these instruments. Without competent examiners, the obtained results may not be a true representation of the examinee's abilities and decisions made upon these results may be in error.

Previous research has investigated administration and scoring errors on the Wechsler Intelligence Scale for Children and the Wechsler Intelligence Scale for Children-Revised (i.e., Sattler, Winget & Roth, 1969; Slate & Jones, 1990). This research has demonstrated that both practitioners and trainees make frequent errors when scoring the WISC and WISC-R. Previous research has also consistently shown that examiner errors most frequently occur on the Vocabulary, Comprehension, and Similarities subtests (Miller & Chansky, 1972; Slate & Jones, 1988; Slate & Chick, 1989). However, research investigating scoring errors on the Wechsler Intelligence Scale for Children-Third Edition is lacking as is research examining scoring differences between graduate

students and practitioners. Scoring errors are particularly important because they have a direct impact on the reliability and validity of obtained test scores. Furthermore, without accurate and valid test results, classification and placement decisions will be adversely effected. Thus, there was a need to examine scoring errors on the Wechsler Intelligence Scale for Children-Third Edition (WISC-III) with a specific focus on the subtests that have consistently been prone to errors (viz., Vocabulary, Comprehension, and Similarities).

The present study examined scoring of the WISC-III Vocabulary, Comprehension, and Similarities subtests with a sample of graduate students and practitioners. Specifically, this study was conducted to determine the extent of scoring errors, if certain response types (Standard/Ambiguous) are more prone to error, and if graduate students and practitioners differ in how they score an identical record form.

While scoring errors were made, the present data tend to conflict with previous research on the WISC and WISC-R (i.e.; Miller, Chansky, & Gredler, 1970; Slate, Jones, Coulter, & Covert, 1992). The current findings suggest that while both graduate students and practitioners make errors when scoring the subtests, these appear to be minimal. However, it should be noted that as many as eight errors occurred on a single subtest (Vocabulary-Ambiguous) and no subtest was without error.

In addition, the present study indicated that graduate students and practitioners do not differ when scoring the same record form. These data also conflict with previous Wechsler Scales research (viz., Slate, Jones, Murray, & Coulter, 1993) where practitioners were more likely to commit errors than graduate students and suggests that

when presented with identical scoring situations, graduate students and practitioners seem to utilize similar criteria when determining the value a given response should receive. This may have resulted because previous research hasn't included both standard and ambiguous items. However, it may also be a direct result of the improved WISC-III *Manual*, which assists examiners by providing a wider variety of scoring examples than previous editions. Furthermore, it is possible that school psychology trainers are spending more time and giving better feedback to graduate trainees on how to correctly score WISC-III items, particularly on the difficult to score subtests (Vocabulary, Comprehension, Similarities) and ambiguous items.

Subtest item type (Standard vs. Ambiguous) in the present study was of significance and similar to prior studies involving both the WISC and WISC-R (Miller, Chansky, & Gredler, 1970; Slate & Chick, 1989; Slate, Jones, Coulter, & Covert, 1992). Regardless of subtest (Vocabulary, Comprehension, Similarities) ambiguous items (items not specifically accounted for in the WISC-III *Manual*) were more prone to error than standard items (items clearly accounted for in the WISC-III *Manual*). Furthermore, these data provide evidence that these scoring errors occur regardless of examiner type (Graduate Student or Practitioner) and suggests that items requiring examiner judgment tend to result in more scoring variability than items specifically assigned a given value in the WISC-III *Manual*.

Based on the current study, it appears that the developers of the WISC-III have made significant improvements over its predecessors in assisting examiners when scoring actual items on the Vocabulary, Comprehension, and Similarities subtests. These

improvements may be a direct result of the expanded scoring items provided in the WISC-III *Manual*. However, scoring errors still occur and future research should continue to examine scorer errors on the WISC-III with more representative samples. Specifically, this research should focus not only on individual subtests but also on an entire record form to determine how scoring errors affect the overall Verbal, Performance, and Full Scale IQ's. Additionally, it would be beneficial for future research to investigate whether or not specific test items can be identified that may be most problematic in scoring. If so, these items may need to be given additional training time, modified, or eliminated from future editions of the WISC-III. Overall, if future research replicates the present findings, users of the WISC-III should obviously continue to follow the standard instructions, scoring guidelines, and required practices of WISC-III administrations but remain particularly aware of ambiguous items that may require significant examiner judgment. Furthermore, trainers of the WISC-III should continue spending adequate time introducing the WISC-III with a focus on ambiguous items and providing substantial practice administrations including proper feedback on scoring errors.

One limitation of the present study is the sample size. Only 21 graduate students from three School Psychology training programs and 23 certified school psychologist practitioners participated in this study. Clearly, these data may have resulted from a small, non-representative group of examiners and caution should be taken before generalizing the findings of this study to a larger, more diverse population of school psychology trainees or practitioners.

A second limitation involved is the examination of only 3 of the 13 WISC-III subtests. Without a completely scored WISC-III protocol, it is impossible to determine how mild or severe scoring errors would affect the overall Verbal, Performance, and Full Scale IQ's which are a significant part of special education evaluations. However, the current data should not be ignored, because it not only indicates that both graduate students and practitioners make errors but that these individuals are also not following standard WISC-III scoring procedures, which suggests potential problems with the reliability and validity of the WISC-III.

References

American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. American Psychologist, 47, 1597-1611.

Bradley, F.O., Hanna, G.S., & Lucas, B.A. (1980). The reliability of scoring the WISC-R. Journal of Consulting and Clinical Psychology, 48, 530-531.

Conner, R., & Woodall, F.E. (1983). The effects of experience and structured feedback on WISC-R error rates made by student-examiners. Psychology in the Schools, 20, 376-379.

Franklin, M.R., Jr., Stillman, P.L., Burpeau, M.Y., & Sabers, D.L. (1982). Examiner error in intelligence testing: Are you a source? Psychology in the Schools, 19, 563-569.

Miller, C.K., & Chansky, N.M. (1972). Psychologists' scoring of WISC protocols. Psychology in the Schools, 9, 144-152.

Miller, C.K., Chansky, N.M., & Gredler, G.R. (1970). Rater agreement on WISC protocols. Psychology in the Schools, 7, 190-193.

Oakland, T., Woo Lee, S., & Axelrad, K.M. (1975). Examiner differences on actual WISC protocols. Journal of School Psychology, 13, 227-233.

Oakland, T., & Zimmerman, S. (1986). The course on individual mental assessment: A national survey of course instructors. Professional School Psychology, 1, 51-59.

Piotrowski, C., & Keller, J. (1989). Psychological testing in outpatient health facilities: A national study. Professional Psychology: research and practice, 20, 423-425.

Reschly, D.J., & Wilson, M.S. (1997). Characteristics of school psychology graduate education: Implications for the entry-level discussion and doctoral-level specialty definition. School Psychology Review, 26, 74-92.

Sattler, J.M., Winget, B.M., & Roth, R.J. (1969). Scoring difficulty of WAIS and WISC Comprehension, Similarities, and Vocabulary responses. Journal of Clinical Psychology, 25, 175-177.

Sherrets, S., Gard, G., & Langner, H. (1979). Frequency of clerical errors on WISC protocols. Psychology in the Schools, 16, 495-496.

Slate, J.R., & Chick, D. (1989). WISC-R examiner errors: Cause for concern. Psychology in the Schools, 26, 78-84.

Slate, J.R., & Jones, C.H. (1988). Sources of examiner errors on the WAIS-R. Social and Behavioral Sciences Documents, 18, 31.

Slate, J.R., & Jones, C.H. (1990). Examiner errors on the WAIS-R: A source of concern. The Journal of Psychology, 124, 343-345.

Slate, J.R., & Jones, C.H. (1990). Identifying students' errors in administering the WAIS-R. Psychology in the Schools, 27, 83-87.

Slate, J.R., & Jones, C.H. (1990). Student error in administering the WISC-R: Identifying problem areas. Measurement and Evaluation in Counseling and Development, 23, 137-140.

Slate, J.R., Jones, C.H., Coulter, C., & Covert, T.L. (1992). Practitioners' administration and scoring of the WISC-R: Evidence that we do err. Journal of School Psychology, 30, 77-82.

Slate, J.R., Jones, C.H., Murray, R.A., & Coulter, C. (1993). Evidence that practitioners err in administering and scoring the WAIS-R. Measurement and Evaluation in Counseling and Development, 25, 156-161.

Stinnett, T.A., Havey, J.M., & Oehler-Stinnett, J. (1994). Current test usage by practicing school psychologists: A national survey. Journal of Psychoeducational Assessment, 12, 331-350.

Warren, S.A., & Brown, W.G., Jr. (1973). Examiner scoring errors on individual intelligence tests. Psychology in the Schools, 10, 118-122.

Whitten, J., Slate, J.R., Jones, C.H. Shine, A.E., & Raggio, D. (1994). Examiner errors in administering and scoring the WPPSI-R. Journal of Psychoeducational Assessment, 12, 49-54.

Appendix A

Demographic Information

Identification Number: _____(Last 4 Digits of Social Security Number)

Gender: _____

Age: _____

Ethnic Background: _____

A) Are you currently enrolled in a graduate program? (Yes / No).

B) If you are currently a practicing school psychologist, please indicate the number of years you have been certified: _____

C) Practitioners-Please indicate the number of case study evaluations you will complete this year: _____

D) Practitioners- Please estimate the percentage of time spent in each of the five following areas: 1) Psychoeducational Assessment _____

2) Direct Intervention _____

3) Problem-Solving Consultation _____

4) Systems Consultation _____

5) Research/Evaluation _____

E) Please indicate the number of Intellectual Assessment Courses completed: _____

F) Please provide an address where you can be contacted in the event your Identification number is selected in the $50.00 drawing:

Appendix B

Instructions

1) Please sign and return the enclosed consent form.

2) Please complete the demographic form provided.

3) Enclosed are copies of the Vocabulary, Comprehension, and Similarities subtests of the WISC-III with responses provided. Please utilize the WISC-III manual and score each response as if it were an actual case. However, please score each item regardless if the discontinue rule for any subtest is met.

4) Please return the scored items in the provided self-addressed stamped envelope.

5) If you would like to receive a copy of the results of this study, please provide a self-addressed mailing label.

Table 1

Percentages of Time Spent in Various Psychological Roles

	Present Sample	Reschly & Wilson (1997)
Psychoeducational Assessment	51.53	56.68
Direct Intervention	23.25	20.25
Problem-Solving Consultation	20.43	16.18
Systems Organization	2.60	5.01
Research/ Evaluation	1.53	1.89

Table 2

Means, Standard Deviations and Ranges of Examiner Errors by Subtest and Item Type

	Student			Practitioner			Total		
	M	SD	Range	M	SD	Range	M	SD	Range
Similarities S	.19	.51	0-2	.13	.34	0-1	.16	.43	0-2
Similarities A	.57	.81	0-2	1.00	.90	0-3	.80	.88	0-3
Comprehension S	.24	.54	0-2	.30	.63	0-2	.27	.59	0-2
Comprehension A	1.76	1.48	0-5	1.83	1.27	0-4	1.80	1.36	0-5
Vocabulary S	.14	.48	0-2	.61	.94	0-3	.39	.78	0-3
Vocabulary A	3.76	1.81	2-8	3.26	1.54	1-7	3.50	1.68	1-8

Table 3

Frequencies of Examiner Errors by Subtest Item Type

Errors	Sim-S	Sim-A	Com-S	Com-A	Voc-S	Voc-A
0	38 (86.4)	21 (47.7)	35 (79.5)	8 (18.2)	33 (75.0)	
1	5 (11.4)	12 (27.3)	6 (13.6)	13 (29.5)	7 (15.9)	2 (4.5)
2	1 (2.3)	10 (22.7)	3 (6.8)	10 (22.7)	2 (4.5)	13 (29.5)
3		1 (2.3)		7 (15.9)	2 (4.5)	11 (25.0)
4				5 (11.4)		6 (13.6)
5				1 (2.3)		8 (18.2)
6						
7						3 (6.8)
8						1 (2.3)

Note. Errors Column = total number of errors committed per subtest, Subtest Columns = number of individuals, Parenthesis = percent of sample. Sim-S = Similarities Standard, Sim-A = Similarities Ambiguous, Com-S = Comprehension Standard, Com-A = Comprehension Ambiguous, Voc-S = Vocabulary Standard, Voc-A = Vocabulary Ambiguous.

Table 4

Analysis of Variance Summary Table for Similarities

	DF	SS	MS	F	p
Between Subjects					
Subject (A)	1	.75	.75	1.55	.220
Error	42	20.21	.48		
Within-Subjects					
Item Type (B)	1	8.58	8.58	19.19	.0001
A X B	1	1.31	1.31	2.93	.094
Error	42	18.78	.45		

Table 5

Analysis of Variance Summary Table for Comprehension

	DF	SS	MS	F	p
Between Subjects					
Subject (A)	1	.09	.09	.06	.808
Error	42	65.30	1.55		
Within-Subjects					
Item Type (B)	1	50.91	50.91	75.05	.0001
A X B	1	.00	.00	.00	.995
Error	42	28.49	.68		

Table 6

Analysis of Variance Summary Table for Vocabulary

	DF	SS	MS	F	p
Between Subjects					
Subject (A)	1	.01	.01	.00	.953
Error	42	81.21	1.93		
Within-Subjects					
Item Type (B)	1	215.86	215.86	148.42	.0001
A X B	1	5.13	5.13	3.53	.067
Error	42	61.08	1.45		